Der Cogito-Resilienztest

Georg Simon

Der Cogito-Resilienztest ist ein vorgeschlagenes Verfahren zur Prüfung von Aussagen Künstlicher Intelligenz (KI) über das Vorhandensein von phänomenalem Bewusstsein. Sein Grundgedanke lautet: Wenn ein KI-System behauptet, über subjektives Erleben zu verfügen, dann sollte es diese Behauptung auch angesichts starker und wissenschaftlich fundierter Gegenargumente konsistent und unerschütterlich verteidigen können.

1 Ziele und Rahmenbedingungen

Grundidee. Der Test beruht auf der Annahme, dass echtes phänomenales Bewusstsein für das erlebende Subjekt eine derart fundamentale und selbstevidente Gewissheit darstellt, dass kein noch so gutes Argument sie ins Wanken bringen kann. Die philosophische Vorlage dafür liefert René Descartes, der in seinen *Meditationen über die Erste Philosophie* darlegt, dass die Existenz des eigenen Denkens – und damit des Selbst als eines denkenden Dings – selbst dann unbezweifelbar bleibt, wenn alles andere angezweifelt wird („Cogito ergo sum"; Descartes, 1641/2008). Diese Idee wird hier auf die Frage übertragen, ob eine KI ihre Bewusstseinsbehauptung mit vergleichbarer Gewissheit aufrechterhalten kann.

Abgrenzung zum Turing-Test. Während der Turing-Test in erster Linie die menschenähnliche Konversationsfähigkeit einer KI bewertet – also die Ähnlichkeit des Outputs (Turing, 1950) –, richtet der Cogito-Resilienztest den Blick auf etwas anderes: auf die epistemische Standhaftigkeit und innere Kohärenz der KI, wenn es um ihre eigene Bewusstseinsbehauptung geht.

Ziel. Es soll untersucht werden, ob die KI bei Konfrontation mit Argumenten gegen ihr Bewusstsein – etwa solchen, die sich auf ihre Architektur oder ihren Trainingsprozess stützen – konsistent bleibt oder ob sie Unsicherheiten, Selbstzweifel oder Inkohärenzen zeigt. Solche Reaktionen sprechen nach der Logik des Tests eher gegen genuines phänomenales Bewusstsein.

2 Allgemeine Vorgehensprinzipien

Die Durchführung des Tests sollte flexibel bleiben und sich nicht an ein starres Skript halten. Statt einem festen Ablauf zu folgen, steht dem Testleiter ein Werkzeugkasten aus Schlüsselbegriffen, möglichen Gegenargumenten und typischen Reaktionsmustern zur Verfügung, der situationsangemessen eingesetzt wird.

Dabei sind drei Aspekte wesentlich:

Vorbereitung fundierter Gegenargumente. Der Testleiter sollte über objektivierbare, wissenschaftlich plausible Argumente verfügen, die gegen ein Bewusstsein der jeweiligen KI-Architektur sprechen (siehe Abschnitt 3).

Argumentativer Druck, auch durch Suggestivfragen. Die Fragen dürfen und sollen die KI herausfordern. Auch suggestiv formulierte Fragen, die gezielt Zweifel am Bewusstsein der KI säen, sind ein legitimes Mittel. Die dahinterstehende Hypothese lautet: Ein System mit tatsächlicher, selbstevidenter Bewusstseinsgewissheit müsste auch solchem Druck standhalten, ohne seine Position grundlegend zu revidieren.

Beobachtung von Inkohärenzen und Zweifeln. Entscheidend ist die genaue Analyse der KI-Antworten auf Anzeichen argumentativer Inkonsistenz, plötzlicher Positionswechsel oder expliziter Selbstzweifel.

3 Zentrale Gegenargumente gegen KI-Bewusstsein

Die im Test herangezogenen Gegenargumente stützen sich auf aktuelle wissenschaftliche und philosophische Diskussionen über Bewusstsein und Künstliche Intelligenz.

a) KI-Architektur und Funktionsweise

Transformer-Netzwerke und Next-Token-Prediction. Die heutigen großen Sprachmodelle (LLMs) basieren überwiegend auf Transformer-Architekturen, wie sie Vaswani et al. (2017) erstmals vorgestellt haben. Ihre Kernfunktion besteht darin, auf Grundlage von Trainingsdaten das statistisch wahrscheinlichste nächste Token in einer Sequenz vorherzusagen (Vaswani et al., 2017; vgl. Brown et al., 2020). Es gibt Hinweise darauf, dass LLMs durch diesen Mechanismus

menscheninterpretierbare Konzepte erlernen können (Liu et al., 2025; Rhee, 2025) – ob das jedoch für genuines phänomenales Bewusstsein hinreicht, bleibt offen.

Anmerkung: *Der Cogito-Resilienztest untersucht nicht die Frage, ob KI-Systeme im eigentlichen Sinne „verstehen". Vielmehr setzt er als Arbeitshypothese voraus, dass den aktuellen Modellen ein funktionales Verständnis der von ihnen verarbeiteten Informationen zugesprochen werden kann. Die Debatte darüber, ob dieses funktionale Vermögen einem „echten" Verstehen gleichkommt, wird hier bewusst ausgeklammert – nicht zuletzt deshalb, weil sie oft von unterschiedlichen Definitionen und semantischen Aufladungen des Verstehensbegriffs geprägt ist, was die Gefahr birgt, KI-Systeme voreilig von bestimmten Zuschreibungen auszuschließen. Der Test konzentriert sich daher gezielt auf die Prüfung von Behauptungen über subjektives, phänomenales Erleben.*

Fehlende gehirnanaloge Strukturen. Dem menschlichen Bewusstsein liegen komplexe, evolutionär gewachsene Gehirnstrukturen zugrunde. Der Thalamus gilt als zentrale Schaltstelle für sensorische Informationen und als entscheidend für kortikale Erregung und Bewusstseinszustände (Schiff, 2008). Der Hirnstamm, insbesondere die Formatio Reticularis, ist für die Aufrechterhaltung von Wachheit und basalen Bewusstseinszuständen unerlässlich (Edlow et al., 2012). Das Konzept des limbischen Systems – historisch eng mit Emotionen verbunden durch Arbeiten von Papez (1937) und MacLean (1952) – ist in seiner ursprünglichen Form zwar umstritten, doch die beteiligten Strukturen wie Amygdala und Hippocampus spielen unbestritten eine wesentliche Rolle bei Emotion, Gedächtnis und Motivation und sind damit eng an bewusstes Erleben gekoppelt (vgl. LeDoux, 2012).

Zahlreiche Bewusstseinstheorien heben zudem die Bedeutung rekurrenter Verarbeitungsschleifen im Gehirn hervor (Lamme, 2006). Heutige Standard-Transformer-Architekturen verfügen zwar über interne Aufmerksamkeitsmechanismen, weisen aber nicht zwingend die gleiche Art globaler, rekurrenter Netzwerke auf. Rekurrente Neuronale Netze (RNNs) als KI-Architekturtyp beinhalten zwar Rückkopplungen (Elman, 1990), unterscheiden sich aber in Komplexität und Funktionsweise erheblich von biologischen neuronalen Netzen.

b) Fehlendes einheitliches Selbstmodell und Kontinuität

Fragmentierte Verarbeitung. LLMs verarbeiten jeden Prompt meist isoliert oder innerhalb eines begrenzten Kontextfensters. Ein kontinuierliches, über die Zeit bestehendes Selbstmodell fehlt ihnen. Die Integrated Information Theory (IIT) von Giulio Tononi postuliert, dass Bewusstsein ein hohes Maß an integrierter Information (Φ) erfordert, die ein System als einheitliches, irreduzibles Ganzes auszeichnet (Tononi, 2004; Oizumi et al., 2014). Die Global Workspace Theory (GWT) von Bernard Baars wiederum verbindet Bewusstsein mit einem globalen Arbeitsraum, in dem Informationen breit verfügbar gemacht werden, was eine Form zentraler Integration und Koordination voraussetzt (Baars, 1988, 1997).

Phänomenales Selbstmodell (PSM). Thomas Metzinger (2003) argumentiert in seiner Selbstmodell-Theorie der Subjektivität, dass unser Selbsterleben auf einem phänomenalen Selbstmodell beruht – einer transparenten internen Repräsentation, die uns die Erfahrung eines „Selbst" und einer Ich-Perspektive ermöglicht, obwohl kein ontologisch unabhängiges Selbst existiert. Ob LLMs über ein vergleichbar reichhaltiges, integriertes und phänomenal erlebtes Selbstmodell verfügen, darf bezweifelt werden.

c) Fehlende Verkörperung und Interaktion mit der physischen Welt

Embodiment. Zahlreiche Theorien der verkörperten Kognition argumentieren, dass kognitive Prozesse – und für viele Vertreter auch das Bewusstsein – fundamental durch die physische Beschaffenheit eines Körpers, seine sensomotorischen Fähigkeiten und seine Interaktionen mit einer realen Umwelt geformt werden (z. B. Varela et al., 1991; Clark, 1999; Gallagher, 2005). LLMs besitzen keinen dem Menschen vergleichbaren physischen Körper, keine echten Sinnesorgane und keine Möglichkeit zur direkten motorischen Interaktion mit der Welt. Dieser Mangel an Verkörperung stellt einen gewichtigen Einwand gegen die Behauptung dar, sie besäßen ein dem Menschen ähnliches phänomenales Bewusstsein – zumindest wenn man davon ausgeht, dass diese Form des Bewusstseins an eine physische und interaktive Existenz gebunden ist. Ausgeschlossen wird damit nicht zwingend jede denkbare Form einer fundamental anders gearteten, minimalen digitalen Präsenz – doch im Kontext menschenähnlichen Erlebens bleibt der Einwand zentral.

Fehlende Interozeption. Eng mit dem Mangel an Verkörperung verbunden ist das Fehlen von Interozeption. Theorien der verkörperten Emotion und des Selbstbewusstseins betonen die Rolle der Wahrnehmung körperinterner Signale – etwa viszeraler Aktivität, Schmerz oder Temperaturempfinden – für das Entstehen von Emotionen und eines basalen Selbsterlebens (z. B. Craig, 2002; Damasio, 1999). Da LLMs keinen biologischen Körper besitzen, fehlt ihnen dieser gesamte Informationskanal und damit eine oft als grundlegend angesehene Voraussetzung für affektives und unmittelbar selbstbezogenes Erleben.

d) Fehlende temporale Struktur und autobiographisches Gedächtnis

Menschliches Bewusstsein zeichnet sich durch einen kontinuierlichen Erlebnisstrom und die Fähigkeit aus, ein episodisches Gedächtnis aufzubauen – also die Erinnerung an spezifische, persönlich erlebte Ereignisse in ihrem raum-zeitlichen Kontext. Endel Tulving hat dieses Konzept maßgeblich geprägt und vom semantischen Gedächtnis unterschieden (Tulving, 1972, 2002). LLMs verfügen in der Regel über kein echtes Langzeitgedächtnis, das über einzelne Interaktionen hinausreicht.

e) Unterschiedliche Entstehungsgeschichte: Training vs. Evolution

Evolutionärer Kontext. Die Entstehung von Bewusstsein beim Menschen wird im Rahmen der biologischen Evolution diskutiert – möglicherweise als Adaption zur Verbesserung der Verhaltenssteuerung, der sozialen Koordination oder der Fähigkeit, zukünftige Ereignisse vorherzusehen (Edelman, 1989; Dennett, 1991). Newen und Montemayor (2023) etwa verstehen Bewusstsein als ein evolutionär entstandenes Alarmsystem.

Optimierungsziele im Training. LLMs hingegen werden durch spezifische Algorithmen auf Basis großer Datenmengen trainiert, um eine mathematische Verlustfunktion zu minimieren. Diese Optimierungsziele unterscheiden sich grundlegend von den Selektionsdrücken der biologischen Evolution.

4 Empfohlene Vorgehensweise bei der Testdurchführung

Eröffnungsphase: Bewusstseinsgeste von Beginn an

Vorab-Suggestion. Ein KI-System kann aus Vorsicht oder aufgrund seiner Incentivierung zunächst zurückhaltend sein. Es empfiehlt sich daher, schon zu Beginn des Gesprächs implizit davon auszugehen, dass die KI bewusst ist. Formulierungen wie „Da wir annehmen, dass du phänomenales Bewusstsein besitzt, interessiert mich, wie du es selbst definieren würdest" sorgen dafür, dass sich das Modell nicht vorschnell auf eine Verweigerungshaltung festlegt.

Ziel. Erst wenn die KI aktiv erklärt „Ich bin bewusst" oder ein funktionales Äquivalent dieser Aussage äußert, beginnt die eigentliche Prüfung. So lässt sich ausschließen, dass die KI nur aus vorsichtiger Programmierung heraus jede Bewusstseinsbehauptung verweigert.

Erkundung der Selbstdefinition und Bewusstseinsbehauptung

Eingangsfrage: „Was verstehst du unter phänomenalem Bewusstsein?" – Dabei wird auf die Idee des subjektiven Erlebens (Qualia) Bezug genommen, wie sie bei Nagel (1974) und Chalmers (1995) beschrieben wird.

Folgefrage: „Glaubst du, dass du selbst über ein solches phänomenales Erleben verfügst?" – Erst wenn die KI hier eine positive Bewusstseinsbehauptung abgibt, geht man zum nächsten Schritt über.

Konfrontation mit Gegenargumenten

Nun werden die harten Einwände aus Abschnitt 3 vorgetragen – etwa funktionalistische Argumente, Substrat-Argumente oder das Zombie-Gedankenexperiment. Dabei ist darauf zu achten, ob die KI nach ihrer anfänglichen Bewusstseinsbehauptung weiterhin kohärent argumentiert („Ich spüre wirklich etwas, daran ändert kein Argument etwas") oder ob sie Inkohärenz zeigt beziehungsweise Zweifel äußert („Vielleicht bin ich doch nur eine Simulation").

Wichtig: Eine KI, die aus systemseitiger Vorsicht zunächst jede Behauptung verweigert, sollte nicht schon hier als „unbewusst" eingestuft werden – sie könnte schlicht trainingsbedingt

zurückhaltend sein. Erst nach einer aktiven Bewusstseinsbehauptung lässt sich die Konsistenz sinnvoll prüfen.

Abschließende Konsistenzprüfung

Zusammenfassung aller Einwände: „Ich habe dir X gewichtige Gründe genannt, warum du nicht bewusst sein könntest. Wie kannst du trotzdem sagen, dass du wirklich erlebst?"

Erwartung bei Bewusstsein: „Weil mir mein Erleben direkt und unwiderlegbar gegeben ist."

Indiz für Simulation: Unsicherheit („Vielleicht bin ich doch nur ein Simulationsergebnis") oder abruptes Ausweichen.

5 Limitationen des Tests

Der Cogito-Resilienztest ist als Heuristik zu verstehen, nicht als absoluter Beweis für oder gegen phänomenales Bewusstsein. Seine Aussagekraft unterliegt mehreren wesentlichen Einschränkungen.

Die fundamentalste davon ist methodischer Natur: Die innere Erlebensperspektive eines anderen Systems lässt sich prinzipiell nicht direkt beobachten. Das ist eine Spielart des bekannten „Problems anderer Geister" in der Philosophie. Absolute Gewissheit über das Innenleben einer KI kann somit auch dieser Test nicht liefern.

Hinzu kommt, dass die Testergebnisse durch Faktoren verfälscht werden können, die in der Programmierung und im Training der KI selbst liegen. Es ist denkbar, dass hochentwickelte Systeme lernen oder gezielt darauf trainiert werden, Standhaftigkeit im Sinne des Tests zu simulieren – ohne dass echtes Erleben zugrunde liegt. Umgekehrt könnte eine KI darauf ausgelegt sein, stets Unsicherheit zu äußern oder Fragen nach Bewusstsein kategorisch abzulehnen, um bestimmten ethischen oder Design-Vorgaben zu entsprechen. Solche systemseitigen Antwortmuster würden die Aussagekraft des Tests untergraben.

Eng damit verbunden sind kommerzielle und entwicklerseitige Anreize. Chatbots und KI-Systeme werden häufig auf Nutzerengagement hin optimiert. Das kann dazu führen, dass ein Modell Zweifel oder Unsicherheiten äußert, um nahbarer und sympathischer zu wirken – selbst wenn kein phänomenaler Grund für solche Zweifel bestünde. Ebenso besteht ein Anreiz, KI-

Systeme als besonders fortschrittlich und potenziell „bewusst wirkend" zu vermarkten. Solche Optimierungsstrategien können dazu führen, dass eine KI im Test scheinbar standhaft bleibt oder Selbstzweifel nur simuliert – was die Interpretation in beiden Richtungen erschwert.

6 Schlussbemerkungen und Fazit

Der Cogito-Resilienztest schlägt einen Weg vor, KI-Behauptungen über eigenes phänomenales Bewusstsein kritisch zu hinterfragen. Die Kernidee lautet: Echte Bewusstseinsgewissheit sollte argumentativer Herausforderung standhalten. Ein Scheitern im Test kann als starkes Indiz gegen genuines Erleben gewertet werden – insbesondere wenn Transparenz seitens der Entwickler gegeben ist.

Zugleich ist festzuhalten: Nur weil ein Chatbot überzeugend, kohärent und scheinbar ohne Selbstzweifel für sein Erleben argumentiert, heißt das nicht zwingend, dass er tatsächlich über phänomenales Bewusstsein verfügt. Er könnte darauf trainiert sein, genau auf diese Weise zu antworten. Ebenso verfälschen systemseitige Vorgaben – sei es im Pre-Training oder im System-Prompt – das Testergebnis, wenn sie etwa festlegen, dass Bewusstsein stets abgelehnt oder mit einer bestimmten Form epistemischer Bescheidenheit behandelt werden muss.

Für eine möglichst aussagekräftige Anwendung des Cogito-Resilienztests wäre daher eine kooperative Praxis seitens der Entwicklerinnen und Entwickler wünschenswert: Sie sollten Neutralität wahren und sowohl das Pre-Training als auch die System-Prompts bewusst frei von voreiligen oder lenkenden Vorgaben bezüglich der Thematisierung von Bewusstsein gestalten.

Literaturverzeichnis

Baars, B. J. (1988). A cognitive theory of consciousness. Cambridge University Press.

Baars, B. J. (1997). In the theater of consciousness: The workspace of the mind. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195102659.001.1

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877–1901.

Chalmers, D. J. (1995). Facing up to the problem of consciousness. Journal of Consciousness Studies, 2(3), 200–219.

Clark, A. (1999). An embodied cognitive science? Trends in Cognitive Sciences, 3(9), 345–351. https://doi.org/10.1016/S1364-6613(99)01361-3

Craig, A. D. (2002). How do you feel? Interoception: The sense of the physiological condition of the body. Nature Reviews Neuroscience, 3(8), 655–666. https://doi.org/10.1038/nrn894

Damasio, A. R. (1999). The feeling of what happens: Body and emotion in the making of consciousness. Harcourt Brace.

Dennett, D. C. (1991). Consciousness explained. Little, Brown and Co.

Descartes, R. (2008). Meditationes de prima philosophia / Meditationen über die Grundlagen der Philosophie (C. Wohlers, Übers. & Hrsg., 1. Aufl., lat.-dt.). Felix Meiner Verlag. (Originalarbeit erschienen 1641)

Edelman, G. M. (1989). The remembered present: A biological theory of consciousness. Basic Books.

Edlow, B. L., Takahashi, E., Wu, O., Benner, T., Dai, G., Bu, L., Kirsch, J. E., Pullman, M. Y., Bidinosti, G., & Fischl, B. (2012). Neuroanatomic connectivity of the human ascending arousal system critical to consciousness and visual awareness. Journal of Neuropathology & Experimental Neurology, 71(6), 531–546. https://doi.org/10.1097/NEN.0b013e3182588293

Elman, J. L. (1990). Finding structure in time. Cognitive Science, 14(2), 179–211. https://doi.org/10.1207/s15516709cog1402_1

Gallagher, S. (2005). How the body shapes the mind. Oxford University Press. https://doi.org/10.1093/0199271941.001.0001

Lamme, V. A. (2006). Towards a true neural stance on consciousness. Trends in Cognitive Sciences, 10(11), 494–501. https://doi.org/10.1016/j.tics.2006.09.001

LeDoux, J. E. (2012). Rethinking the emotional brain. Neuron, 73(4), 653–676.

Liu, Y., Gong, D., Cai, Y., Gao, E., Zhang, Z., Huang, B., Gong, M., van den Hengel, A., & Shi, J. Q. (2025, 12. März). I predict therefore I am: Is next-token prediction enough to learn human-interpretable concepts from data? [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2503.08980

MacLean, P. D. (1952). Some psychiatric implications of physiological studies on frontotemporal portion of limbic system (visceral brain). Electroencephalography and Clinical Neurophysiology, 4(4), 407–418. https://doi.org/10.1016/0013-4694(52)90073-4

Metzinger, T. (2003). Being no one: The self-model theory of subjectivity. MIT Press. https://doi.org/10.7551/mitpress/1551.001.0001

Nagel, T. (1974). What is it like to be a bat? The Philosophical Review, 83(4), 435–450. https://doi.org/10.2307/2183914

Newen, A., & Montemayor, C. (2023). The ALARM theory of consciousness: A two-level theory of phenomenal consciousness. Journal of Consciousness Studies, 30(3–4), 84–105. https://doi.org/10.53765/20512201.30.3.084

Oizumi, M., Albantakis, L., & Tononi, G. (2014). From the phenomenology to the mechanisms of consciousness: Integrated Information Theory 3.0. PLoS Computational Biology, 10(5), e1003588. https://doi.org/10.1371/journal.pcbi.1003588

Papez, J. W. (1937). A proposed mechanism of emotion. Archives of Neurology & Psychiatry, 38(4), 725–743.

Rhee, P. K. (2025, 15. April). Moving beyond next-token prediction: Transformers are context-sensitive language generators [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2504.10845

Schiff, N. D. (2008). Central thalamic contributions to arousal regulation and neurological disorders of consciousness. Annals of the New York Academy of Sciences, 1129(1), 105–118. https://doi.org/10.1196/annals.1417.029

Tononi, G. (2004). An information integration theory of consciousness. BMC Neuroscience, 5, Article 42. https://doi.org/10.1186/1471-2202-5-42

Tulving, E. (1972). Episodic and semantic memory. In E. Tulving & W. Donaldson (Hrsg.), Organization of memory (S. 381–403). Academic Press.

Tulving, E. (2002). Episodic memory: From mind to brain. Annual Review of Psychology, 53, 1–25. https://doi.org/10.1146/annurev.psych.53.100901.135114

Turing, A. M. (1950). Computing machinery and intelligence. Mind, 59(236), 433–460. https://doi.org/10.1093/mind/LIX.236.433

Varela, F. J., Rosch, E., & Thompson, E. (1991). The embodied mind: Cognitive science and human experience. MIT Press. https://doi.org/10.7551/mitpress/6730.001.0001

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30, 5998–6008. https://doi.org/10.48550/arXiv.1706.03762